Multiple Fundamental Frequency Estimation and Polyphony Inference of Polyphonic Music Signals

Chunghsin Yeh, *Member, IEEE*, Axel Roebel, *Member, IEEE*, and Xavier Rodet, *Member, IEEE*

***Abstract*—This paper presents a frame-based system for estimating multiple fundamental frequencies (F0s) of polyphonic music signals based on the short-time Fourier transform (STFT) representation. To estimate the number of sources along with their F0s, it is proposed to estimate the noise level beforehand and then jointly evaluate all the possible combinations among pre-selected F0 candidates. Given a set of F0 hypotheses, their hypothetical partial sequences are derived, taking into account where partial overlap may occur. A score function is used to select the plausible sets of F0 hypotheses. To infer the best combination, hypothetical sources are progressively combined and iteratively verified. A hypothetical source is considered valid if it either explains more energy than the noise, or improves significantly the envelope smoothness once the overlapping partials are treated. The proposed system has been submitted to Music Information Retrieval Evaluation eXchange (MIREX) 2007 and 2008 contests where the accuracy has been evaluated with respect to the number of sources inferred and the precision of the F0s estimated. The encouraging results demonstrate its competitive performance among the state-of-the-art methods.**

***Index Terms*—Automatic music transcription, frequency estimation, music information retrieval, noise estimation, signal analysis, source separation.**

I. Introduction

FUNDAMENTAL frequency, or F0, is an essential descriptor of harmonic sound signals such as speech and music. Single-F0 estimation algorithms assume that there is at most one harmonic source of which the F0 is to be extracted. Although single-F0 estimation algorithms have been considerably developed, their applications to music signals are somehow limited because most music signals contain several concurrent harmonic sources.[1] Multiple-F0 estimation algorithms are thus required for the general case and the estimation of the number of sources, called *polyphony inference*, has to be addressed. The problem of estimating the F0s of a monodic instrument solo recording is already challenging. Musical instruments produce sounds with various spectral envelopes, inharmonic partials, and spurious components [1], [2]. When reverberation is involved, it prolongs preceding sound events such that they overlap with the following events, giving rise to polyphonic signals [3], [4]. When there are several musical instruments playing, sound source components may overlap in time and frequency, resulting in more complex sound mixtures. The complexity causes the octave ambiguity as well as the source number ambiguity.

In general, polyphonic signals consisting of multiple harmonic sound sources can be expressed as

$$y[n] = \sum_{m=1}^{M} y_m[n] + z[n] \tag{1}$$

where n is the discrete time index, M is the number of harmonic sources (the polyphony), $y_m[n]$ is the quasi-periodic part of the mth source and $z[n]$ is the noise part. The main problem to deal with is the modeling of $y_m[n]$ and the decomposition of the observed signal into an unknown number of model sources, which is actually a problem of pattern matching. Despite a variety of existing methods, they mainly follow three principles that are derived from the physical properties of musical instrument sounds: (1) harmonicity, (2) spectral smoothness [5], and (3) synchronous amplitude evolution within a single source [6]. These physical properties are closely related to certain perceptual mechanisms that the human auditory system uses to segregate harmonically related spectral components forming a smooth envelope ([7, p. 232]) and having a similar temporal evolution ([7, p. 575]). Some approach this problem under a global mathematical (formulation/optimization) scheme, such as statistical adaptation of waveform models [8], [9], non-negative matrix factorization based methods [10]–[12], "specmurt" analysis [13], harmonic temporal structured clustering (HTC) method [14]. Others deal with each subproblem independently [5], [6], [15]–[17] and a probabilistic framework like hidden Markov model is often used for tracking [18]–[21]. Extensive study of the problem of multiple-F0 estimation leads us to conclude that there are three fundamental models involved. 1) *Source model*: The use of harmonic patterns to match the observed signal is the common technique for most of the existing methods. Owing to the complex nature of musical instrument sounds, it is necessary to adapt the harmonic patterns of $y_m[n]$ to a variety of time-varying instrument sounds. 2) *Source interaction model*: Since the partials of harmonic sources may overlap, it is important to handle this situation such that a combination of harmonic patterns can be more reasonably matched to the observed signal. For simplicity, many existing methods assume the additivity of linear/power spectrum for the overlap of partials [22]. 3) *Noise model*: Little attention is drawn to noise estimation. However, we believe that an explicit noise modeling provides an appropriate means for estimating the

Manuscript received September 16, 2008; revised July 12, 2009. First published August 11, 2009; current version published July 14, 2010. the MRNT (le Ministère déléguéà la Recherche et aux Nouvelles Technologies) of France as part of the MusicDiscover project (2004–2007). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sylvain Marchand.

The authors are with the Institut de Recherche et Coordination Acoustique/Musique (IRCAM), 75004 Paris, France (e-mail: chunghsin.yeh@ircam.fr).

Digital Object Identifier 10.1109/TASL.2009.2030006

[1]A note played by a musical instrument is considered a harmonic sound source.

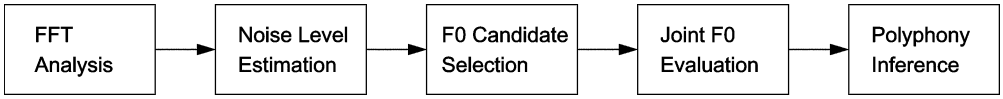


Fig. 1. Overview of the proposed multiple-F0 estimation system. Noise level estimation serves to distinguish the sinusoids from the noise in the observed spectrum given by the fast Fourier transform (FFT) analysis. After selecting the F0 candidates, their possible combinations are jointly evaluated and the number of sources is estimated by a polyphony inference algorithm.

number of sources M, whereas many of the existing methods use heuristic thresholds related to the source energy or the residual energy. The major drawback is that these thresholds are not adaptive for different conditions of signal-to-noise ratios. Others model the noise as white Gaussian noise [8], [9] which is not appropriate for the colored noise that is usually observed.

Following the physical/perceptual principles, we aim at treating these key problems by proposing the algorithms for noise estimation, harmonic matching adapted for polyphonic music signals and overlapping partial treatment. Another well-known problem of F0 estimation is the prevention of subharmonic/super-harmonic errors which is very challenging while dealing with polyphonic signals. For example, common subharmonics having the support from the partials of concurrent sources can compete with the correct F0s. Therefore, complementary criteria are proposed to prevent these errors. Instead of using heuristic thresholds, we propose to refer to monophonic musical instrument sounds for deriving coherent thresholds.

This paper is organized as follows. In Section II, the overview of the proposed multiple-F0 estimation method is given. Then, the algorithms in each part of the system are presented: noise level estimation (Section III), joint evaluation of F0 hypotheses (Section IV), candidate selection (Section V), and polyphony inference (Section VI). In Section VII, the evaluation results are presented, including the Music Information Retrieval Evaluation eXchange (MIREX) 2008 and 2009 results. Finally, conclusions are drawn and perspectives are discussed.

II. System Overview

A frame-based multiple-F0 estimation system for single-channel polyphonic music signals is to be presented (see Fig. 1). Such a system can later include temporal information by means of a tracking mechanism to build continuous F0 trajectories, which can be used for automatic music transcription, source separation, etc. The algorithms developed in the system are based on the sinusoids plus noise signal model of which the components are derived from spectral peaks [23], [24]. Because the spectral peaks are representative of the components generated from the harmonic sources, they give a direct access to analyze the underling sources. The system starts with adaptive noise level estimation which classifies the spectral peaks into sinusoids and noise. The sinusoidal peaks are considered partials of harmonic sources that a combination of harmonic patterns related to F0 hypotheses will match. To jointly evaluate a combination of hypothetical sources, their partials are estimated by harmonic matching and the overlapping partials are treated. Then, four criteria are used together to score the plausibility of a hypothetical combination.

TABLE I
Models and Assumptions of the Proposed Method

observed signal	represented as successive spectral peaks classified as sinusoid/noise
noise model	frequency-dependent envelope of limited cepstral order applied to white noise
source model	quasi-harmonic model with smooth spectral envelopes
interaction model	the amplitude of overlapping partials determined by the strongest source

The polyphony hypothesis is progressively increased and all possible combinations are evaluated. A candidate selection method is also proposed such that the number of combinations to evaluate is reasonably reduced. Finally, the most plausible combination is determined by a polyphony inference algorithm. The model assumptions made are listed in Table I.

III. Adaptive Noise Level Estimation

Contrary to most methods that do not explicitly model the noise part of the signal, a probabilistic description of the noise level is proposed. If the noise part is not estimated beforehand, the number of sources can be overestimated when unnecessary sources simply explain the noise. In a previous study the statistical properties of the deterministic signals embedded in white Gaussian noise have been derived and used to distinguish deterministic components from noise in the spectrogram [25]. It makes use of the chi-squared distribution to model the energy spectrum of noisy signals. To be able to estimate the time-varying colored noise, we take a different approach which makes use of the Rayleigh distribution to model the spectral magnitude distribution of noise [26]. The noise is understood as generated from white noise filtered by a frequency-dependent spectral envelope; the noise level is defined as the expected magnitude level of noise peaks. It is assumed that the noise level is slowly varying with frequency such that it can be modeled by means of a Rayleigh distribution with frequency-dependent mode. A Rayleigh random variable X has probability density function [27]:

$$p(x) = \frac{x}{\sigma^2} e^{-x^2/(2\sigma^2)} \text{ with } 0 \leqslant x < \infty, \quad \sigma > 0. \tag{2}$$

Consider the Rayleigh random variable X as the observed magnitudes of noise peaks in a narrow band, then the *Rayleigh mode* σ specifies the most frequently observed magnitudes of noise peaks. The expected value of the noise spectrum is given by

$$E[X] = \sigma\sqrt{\frac{\pi}{2}} \tag{3}$$

which is referred to as (*mean*) *noise level*. The pth percentile

$$x_p = \sigma\sqrt{-2\log(1-p)}, \quad 0 < p < 1 \tag{4}$$

is referred to as *noise envelope* which provides a threshold that can be used to classify the sinusoidal peaks in the spectrum, given the Rayleigh mode σ and a user selected control parameter p that determines the probability of misclassification.

The main problem here is to establish an estimate of the noise level given only the observed spectrum. Using the spectral peak descriptors presented in [28], the sinusoids obeying the allowed amplitude/frequency modulation rates can be detected [29]. The sinusoids are then removed from the spectrum. The residual spectrum is expected to contain the noise as well as the sinusoids with larger modulation rates. For example, colliding sinusoids that represent the overlapping partials of concurrent harmonic sources. These peaks can further be distinguished by means of the statistical properties of the Rayleigh distribution. To this end, an iterative approximation process of the noise level is proposed. According to the relation between the Rayleigh mode and the expected value of the log-amplitude spectrum [30]:

$$E[\log X] \approx \log(\sigma) + 0.058 \tag{5}$$

the frequency-dependent expected amplitude of the residual spectrum can be derived using (3). Notice that $E[\log X]$ can be estimated by means of cepstral liftering [31], [32]. In each iteration, the residual amplitudes of the spectrum are first normalized by the current estimate of the Rayleigh mode and the spectrum is then tested against the hypothesis that it follows a Rayleigh distribution. To achieve an efficient test that can easily be evaluated in the iterative procedure, the test is based on statistical measures and makes use of the fact that mean and variance of the Rayleigh distribution are coupled through the Rayleigh mode. A straightforward approach is to use skewness or kurtosis of which the definition implies both mean and variance. Skewness, defined as the third moment about the mean divided by the third power of the standard deviation [33], is chosen. To increase the speed of the iterative procedure the normalized spectrum is divided into subbands of equal bandwidth[2] and in each subband the skewness is calculated and compared to the skewness of the Rayleigh distribution

$$Skw_{rayl} = \frac{2(\pi-3)\sqrt{\pi}}{\sqrt{(4-\pi)^3}} \approxeq 0.6311 \tag{6}$$

If the observed skewness falls between 0 and Skw_{rayl}, the distribution fit is considered achieved. Otherwise, the largest outlier in the related subband is reclassified as sinusoid and the noise level is updated accordingly. Since only a few sinusoids are expected to remain in the residual spectrum, they initially give rise to a negative skewness. As the outliers are iteratively reclassified, the skewness is expected to increase toward that of the Rayleigh distribution. When all the subbands achieve the distribution fit, the noise level and the noise envelope are obtained (see Fig. 2). The peaks above the noise envelope are classified as sinusoids and those below it are classified as noise. Notice that if the underlying noise level varies significantly within a subband, the procedure may not converge to a reasonable estimate.

[2]For the analysis frequency up to 8 kHz, 25 subbands are used empirically.

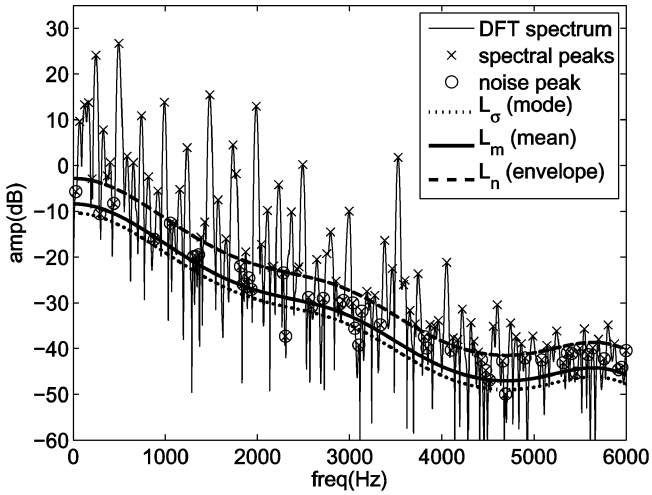


Fig. 2. Estimated noise level for a polyphonic signal. The estimated noise envelope L_n serves as the noise threshold which is user-adjustable. For comparisons, the estimated Rayleigh mode L_σ and the noise level L_m are also shown.

IV. Joint Evaluation of Multiple F0 Hypotheses

Noise level estimation provides a probabilistic classification of the spectral peaks into sinusoids, considered the partials of harmonic sources, and noise. Accordingly, a set of hypothetical sources shall match as many sinusoidal peaks as possible. It is proposed to jointly evaluate the plausibility of a set of F0 hypotheses, following the three physical/perceptual principles. In this section, the description of the joint estimation algorithm is focused on the case in which the number of sources is given. The polyphony inference algorithm will later make use of the jointly evaluated F0 hypotheses to estimate the number of sources.

A. Hypothetical Partial Sequences

Given a set of F0 hypotheses, the frequencies and the amplitudes of their *hypothetical partial sequences* (HPS) are estimated by two processes: 1) partial selection and 2) overlapping partial treatment. The partial selection technique is based on harmonic matching which is generally used for single-F0 estimation [34]. It makes use of a spectral comb with a regular interval of F0 to match the observed spectral peaks. This technique is adopted here with several refinements to construct hypothetical sources, such as inharmonic partial adaptation (see Appendix A). To remove the ambiguity in the overlapping partials of HPS, it is proposed to re-estimate, for each hypothetical source that overlaps, the partial amplitudes based on the interpolation of non-overlapping partials (see Fig. 3 and Appendix B)

B. Score Function

Following the guiding principles, we concentrate on designing four score criteria to evaluate the plausibility of hypothetical sources: harmonicity (HAR), mean bandwidth (MBW), spectral centroid (SPC), and the standard deviation of mean time (SYNC). Conceptually, each criterion is designed to disfavor the F0 hypotheses that are either lower or higher than the correct F0 according to the respective principle. The idea is to have them work in a complementary way such that the subharmonic/super-harmonic errors are prevented.

1) Harmonicity: The score criterion HAR evaluates the harmonic matching between the combination of the hypothetical sources and the observed spectral peaks. To derive the harmonic

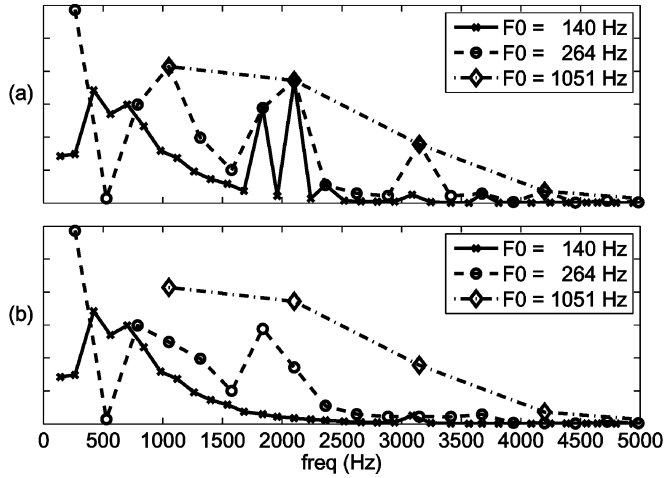


Fig. 3. Hypothetical overlap treatment. (a) HPS constructed by partial selection. (b) HPS after the treatment of the overlapping partials. The lines illustrate the envelopes of the HPS.

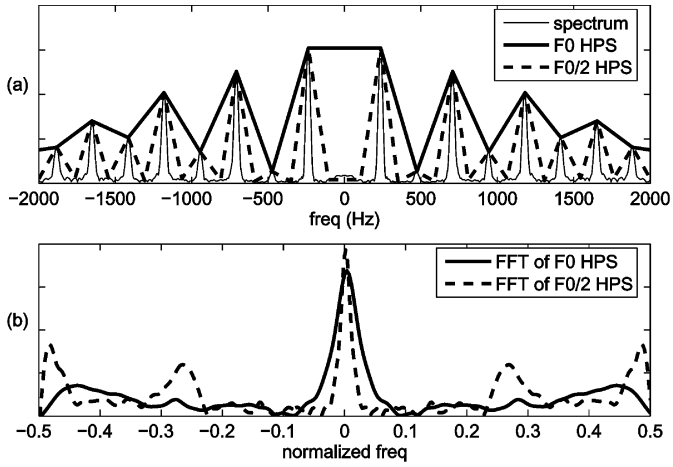


Fig. 4. Spectral smoothness comparison between the HPS of the correct F0 and that of the subharmonic F0/2. The demonstrated sample is a clarinet sound playing the note Bb3.

matching using M hypothetical sources, their individual deviation vectors $d_m(i)$ [see (14)] are first combined as follows:

$$D_M(i) = \min\left(\{d_m(i)\}_{m=1}^{M}\right), \quad \forall\, i \in I \tag{7}$$

where I is the number of the peaks. That is, each observed peak is matched with the closest partial among all HPS such that the resulting combination explains the observed spectrum with the lowest inharmonicity. HAR is then defined as the weighted sum of $D_M(i)$ for all peaks

$$\text{HAR} = \frac{\sum_{i=1}^{I} \text{Spec}(i) \cdot D_M(i)}{\sum_{i=1}^{I} \text{Spec}(i)} \tag{8}$$

where the *peak salience*, $\text{Spec}(i)$, is the square-root of the sum of linear amplitudes for all the bins within the ith spectral peak. The reason of not using the peak energy (the sum of squared amplitudes) is to not emphasize the dynamics of partial amplitudes. It is well known that harmonic matching alone is not adequate for determining the best F0s because the subharmonics will have competitive matching score. Therefore, the following three criteria are designed to compensate HAR.

2) Mean Bandwidth: To score the spectral smoothness of a hypothetical source, the frequency content of the envelope of a HPS is evaluated by means of its bandwidth. By assembling the HPS with its mirrored sequence, a symmetrical sequence g_m with smooth transition in the middle is obtained [see Fig. 4(a)]. Applying K-point[3] FFT to g_m, the related spectrum G_m is acquired [see Fig. 4(b)]. Mean bandwidth is then defined as follows:

$$\text{MBW}_m = \frac{1}{\frac{K}{2}} \sqrt{2 \cdot \frac{\sum_{k=1}^{K/2} k|G_m(k)|^2}{\sum_{k=1}^{K/2} |G_m(k)|^2}} \tag{9}$$

which indicates the degree of energy spread in the high frequency. In this way, the envelope of g_m with smaller variations results in a smaller MBW_m. The function of MBW is to discriminate a correct F0 from its subharmonics. To further illustrate the function of MBW, a clarinet sound signal is used to demonstrate the difference in the envelope smoothness for the HPS of F0 and the HPS of F0/2 (see Fig. 4). The resonance structure of the clarinet sound does not result in a smooth spectral envelope. Nevertheless, the envelope of F0/2 is less smooth than that of F0. Compared with the HPS of F0, the HPS of a subharmonic like F0/2 has higher frequency energy. That is, the energy spreads more widely in frequency and MBW is larger.

3) Spectral Centroid: For harmonic instrument sounds, the spectral centroid tends to lie around the lower partials because the higher partials often decay gradually. According to this general property related to the spectral smoothness principle, the centroid can evaluate the energy spread of a HPS:

$$\text{SPC}_m = \frac{1}{\frac{B}{2}} \sqrt{2 \cdot \frac{\sum_{n=1}^{N_m} n[\text{HPS}_m(n)]^2}{\sum_{n=1}^{N_m} [\text{HPS}_m(n)]^2}} \tag{10}$$

where N_m is the length of HPS_m. B is a normalization factor determined by $\lfloor F_{90}/F0_{\min} \rfloor$. The *spectral roll-off* F_{90} stands for the frequency limit containing 90% of spectral energy in the analysis frequency range [35]. $F0_{\min}$ is the minimal F0 hypothesis in search. Spectral centroid is designed to prevent especially common subharmonic errors. A common subharmonic may relate the partials of several sources to form a smooth envelope, which is favored by MBW. However, SPC tends to disfavor it because the related partials often spread rather widely in frequency.

4) Synchronicity: To evaluate the synchronicity of the temporal evolution of the partials in a HPS, *mean time* is estimated for individual spectral peaks. Mean time is an indication of the center of gravity of the signal energy [36]. It can be defined in the frequency domain as the weighted sum of group delays. The mean time of a spectral peak can be estimated by considering only the frequency bins within a spectral peak, which can characterize the amplitude evolution of the related source [37]. For a coherent HPS, the synchronous evolution of partials is expected,

[3]Two times the next power of 2 of the length of g_m.

which results in a small variance of mean time w.r.t. the matched peaks. The *mean time of a hypothetical source*, denoted by T_m, is calculated as the power spectrum weighted sum of the mean time of the hypothetical partials. The standard deviation of the mean time of the partials is then formulated as

$$\mathrm{SYNC}_m = \frac{1}{\frac{L}{2}} \sqrt{\sum_{i \in \mathrm{HPS}_m} \{(\bar{t}_i - T_m)^2 \cdot w_m(i)\}} \qquad (11)$$

where L is the window size, $\bar{t}_i$ denotes the mean time of the ith observed peak. The weighting vector w_m is constructed from the amplitudes of the HPS. The weight of overlapping partials are set to zero because the spectral phases are possibly disturbed. Since this criterion in some way makes use of the randomness of noise to disfavor an incoherent HPS, an exponential compression factor of 0.23 is applied to w_m in order to raise the significance of the noise components (see the specific loudness descriptor in [35]). In this way, w_m avoids the use of the disturbed phases of overlapping partials, and makes use of the spurious peaks to penalize a HPS matching more noise peaks. w_m is then normalized such that its sum is one.

Notice that the three criteria presented so far are calculated individually for each hypothetical source. To combine the individual criteria into combinatorial ones (MBW, SPC, and SYNC), they are weighted by the *effective salience* of the respective hypothetical sources. The effective salience is the sum of the peak salience of the related partials. The term "effective" is used because the ambiguous partials have been treated such that the impact of the other sound sources on the related partials are at least partially removed. The score function is then formulated as a linear combination of the four criteria:

$$\mathcal{S} = p_1 \cdot \mathrm{HAR} + p_2 \cdot MBW + p_3 \cdot \mathrm{SPC} + p_4 \cdot \mathrm{SYNC} \qquad (12)$$

where $\{p_j\}_{j=1}^4$ are the weighting parameters. Note that the individual score criteria are nonlinear functions of the observed spectra that have been carefully designed to achieve optimal performance (spectral compression, overlap treatment, etc.). Various implementations of the different criteria have been tested and only the best set of criteria is presented here. While a nonlinear combination of the criteria would certainly improve the final result it would complicate the understanding of the score function.

The four criteria are designed in a way that a smaller weighted sum stands for a better score. HAR will slightly favor subharmonic F0s but strongly disfavor super-harmonic F0s; whereas MBW, SPC and SYNC strongly disfavor subharmonic F0s and slightly favor super-harmonic F0s, making use of the respective features of the signal. The weighting parameters are trained to balance the relative support of each criterion such that the score function generally ranks the correct combination on top. The overall scoring mechanism thus remains easy to comprehend.

C. Evaluation for the Case When the Polyphony is Given

Following the evaluation scheme of [5], musical instrument sound samples are semi-randomly mixed with equal energy to create the evaluation database for the case when the polyphony is given [6]. To train the weighting parameters $\{p_j\}_{j=1}^4$, 100 polyphonic samples for each polyphony from one to five are created as the training database. The weighting parameters in the score function are trained by the evolutionary algorithm [38] and the parameter set resulting in the best performance[4] is selected for the evaluation. The joint estimation algorithm is tested for the polyphony from one to five (see Fig. 5). The analysis window size is 93 ms and a correct estimate shall not deviate from the ground truth by more than 3% (a quarter-tone range). To investigate how effectively MBW, SPC, and SYNC compensate HAR, a further test is carried out in which one of the three criteria is deactivated. It is observed that the deactivation of any of the three criteria degrades the overall performance. The result also demonstrates the competitive performance of the proposed algorithm compared to several algorithms mentioned in [39] that are evaluated under a similar scheme.

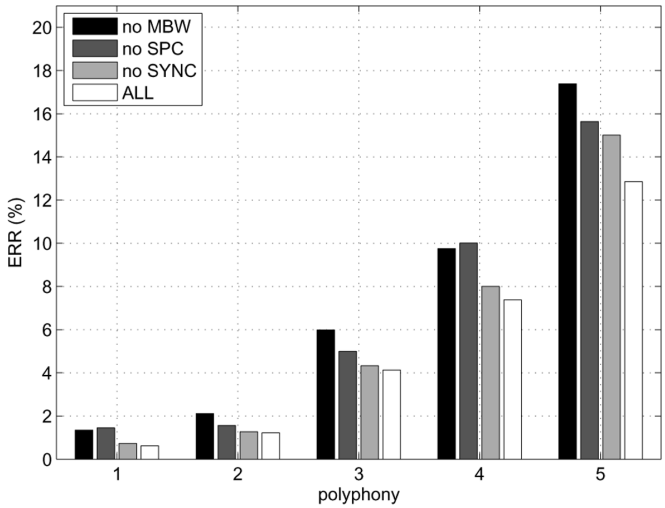


Fig. 5. Evaluation of the score function for the case when the polyphony is given. The functioning of the score criteria are compared: deactivation of MBW (no MBW), deactivation of SPC (no SPC), deactivation of SYNC (no SYNC) and activation of all the criteria (ALL).

V. Candidate Selection

The joint estimation algorithm has a computational concern that the number of combinations grows exponentially with the number of F0 candidates as well as the polyphony. If the F0 candidates are, for instance, sampled on a 1-Hz grid between 50 Hz and 2000 Hz, there will be more than one billion combinations to evaluate for a polyphony of three. A proper candidate selection helps to reduce unnecessary calculations while keeping the robustness of an F0 estimation algorithm. In this section, a candidate selection method is presented. The underlying F0s are seen as two groups. For the F0s that are multiples of another F0, they are harmonically related F0s (HRF0s). Otherwise, they are non-harmonically related F0s (NHRF0s). The partials related to a HRF0 are very likely to be completely overlapped with other sources (e.g., the F0 of 1051 Hz in Fig. 3); whereas those related to a NHRF0 are only partly overlapped (e.g., the F0 of 140 Hz and 264 Hz in Fig. 3). Following this concept, the candidate selection method first extracts the set of NHRF0 sources that match most of the sinusoidal peaks, followed by detecting probable HRF0 sources within the NHRF0 sources.

[4] $\{p_j\}_{j=1}^4 = \{0.3774, 0.2075, 0.2075, 0.2075\}$

A. Extraction of Non-Harmonically Related F0 Candidates (NHRF0s)

The extraction of NHRF0s involves three parts: predominant-F0 estimation, the verification of an extracted F0 candidate and a criterion to stop the iteration. For predominant-F0 estimation, the score function is used as a single-F0 estimator to extract the most probable F0. To suppress an extracted source, the peak salience related to its partials are set to zero. To avoid the extraction of spurious candidates, the *harmonic-to-noise ratio* [32] related to the predominant F0 is evaluated. The overlapping partials are treated beforehand for a less ambiguous evaluation. Similarly, the *residual-to-noise ratio* is calculated for all the peaks that are not yet explained by the F0 candidates. It is meant to indicate if any NHRF0 sources may remain. Accordingly, the extraction process can be terminated when the residual-to-noise ratio falls below a predefined threshold. The F0-dependant thresholds for both ratios are trained on the periodic parts and the noise parts of musical instrument sound samples [40], respectively.

B. Detection of Harmonically Related F0 Candidates (HRF0s)

Each NHRF0 represents a harmonic group within which HRF0s are to be extracted. It is assumed that as long as a HRF0 source is dominant and disturbs the envelope smoothness of the related NHRF0 source, it is reasonable to consider the HRF0 to be an F0 candidate. The same concept is proposed in [15] and the interpolated amplitudes are used as the reference envelope to measure how much the smoothness is disturbed. The issue of this method is that the overlapping partials may be used for interpolation. Since there are no means to locate the overlapping partials in this stage, we propose to refer to the tone models of musical instrument sounds.

Using a collection of samples from McGill University Master Samples, Iowa University Musical Instrument Samples, IRCAM Studio On Line and RWC Musical Instrument Sound Database [41], we group the observed signals according to two types of tone models: *strong-fundamental model* and *weak-fundamental model*. The strong-fundamental model is of a strong fundamental, which represents a spectral envelope with a fast decay for higher partials (see Fig. 6(b)). This corresponds to the general pattern that is used in several existing methods [13], [14], [17]. The weak-fundamental model is of a weak fundamental, which represents a spectral envelope with boosted partials at resonance frequencies higher than the first partial [see Fig. 6(a)]. In each group, the F0-dependent tone models are trained over all instruments for musical notes ranging from Ab1 to B6.

Given the HPS of a NHRF0, the matched tone model is selected according to the least squared error. The partials exceeding the envelope of the tone model are possibly generated from one or more HRF0 sources within the NHRF0 source. Each partial position of the NHRF0 source is considered a HRF0 hypothesis and if it relates to a significant amount of the exceeding partials, it is extracted as a HRF0 candidate. To train the threshold for the exceeding partials, we refer again to the musical instrument sound samples and derive the appropriate threshold. The threshold has been derived for each note and for each partial position, averaging over all the instruments [40].

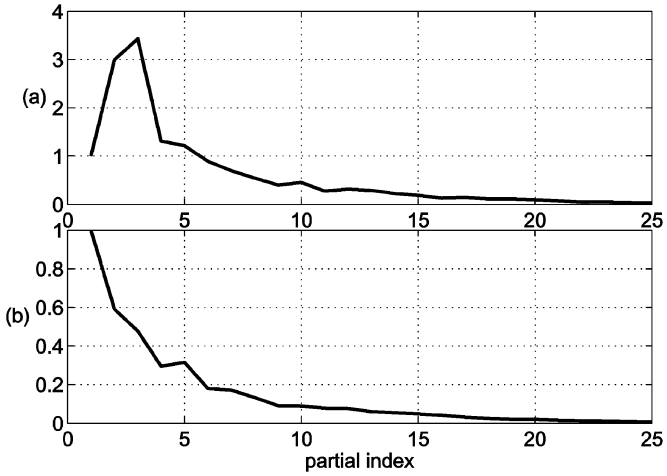


Fig. 6. Two types of tone models for the note E3. (a) Weak-fundamental model and (b) strong-fundamental model. The x-axis represents the partial index; the y-axis represents the relative amplitude.

VI. Polyphony Inference

The estimation of the number of sources is a critical problem of multiple-F0 estimation. Our strategy is to progressively increase the polyphony hypothesis M and calculate the score of all possible combinations of F0 candidates. The scoring of hypothetical combinations is used to select the most plausible ones, among which the best combination is determined by iteratively verifying the related F0 hypotheses to consolidate the estimates. The estimation of the largest polyphony possible $N_{\max}$ relies on the *score improvement* [4]. All the top-five combinations (ranked by the score function), denoted by $\{\mathcal{C}_m\}_{m=1}^{N_{\max}}$, are retained for the consolidation of the final F0 estimates, denoted by $\mathcal{F}$.

The inference algorithm begins with ranking the individual F0 hypotheses found in $\mathcal{C}_{N_{\max}}$, denoted by $\mathcal{H}$, in order of their salience which is derived from the individual score weighted by the appearing "frequency" in $\mathcal{C}_{N_{\max}}$. Beginning with the most salient F0 hypothesis, each hypothesis is consecutively combined with the current estimate $\mathcal{F}$ and verified according to the following criteria.

1) An Additional NHRF0 Source Shall Explain More Than Noise: An added NHRF0 source is considered valid if the reduction of the *residual salience* ΔE_R is larger than the *noise salience* E_{noise}. Both salience is calculated by summing the peak salience of the respective residual or noise peaks. That is, adding a NHRF0 source is reasonable as long as the non-overlapping partials explain a significant amount of salient peaks.

However, adding a source of HRF0 may not reflect a significant ΔE_R if most of the partials are overlapped with other sources. The second criterion is therefore proposed to further validate the HRF0s.

2) Additional HRF0 Source Shall Improve the Spectral Smoothness: Adding a HRF0 source usually improves the smoothness of the spectral envelopes of the previously selected sources. However, a constraint is necessary to prevent adding spurious HRF0s. To achieve this goal, it is proposed to derive the constraint from the score criterion MBW of musical instrument sounds. Given a harmonic sound, each partial frequency

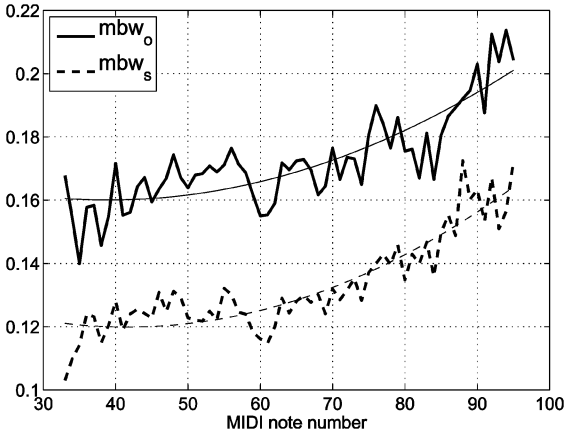


Fig. 7. Comparison of MBW calculated from musical instrument sounds. MBW of the original spectral envelopes. mbw_o and MBW of the smoothed spectral envelopes mbw_s. The two thin curves are second-order polynomial functions fitting the trained MBW data.

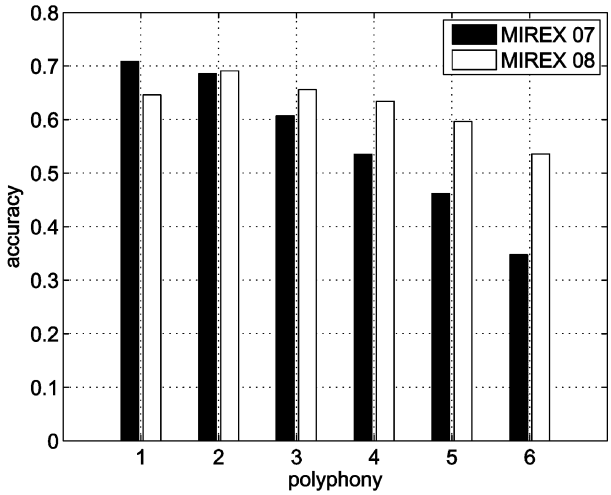


Fig. 8. Comparison of the accuracy rates between the MIREX'07 version and the presented version (MIREX'08). Both are evaluated using 26 pieces of synthesized polyphonic music.

is considered a HRF0 hypothesis. For each HRF0 hypothesis, the decrease of MBW (Δ MBW) is calculated, which is the difference of MBW before, denoted by mbw_o, and after, denoted by mbw_s, smoothing out[5] the related partials. For each analysis instance, mbw_o of the correct F0 and mbw_s of the HRF0 hypothesis that results in the maximal Δ MBW are retained. For each musical note, the calculated mbw_s and mbw_o are averaged for all the analysis instances of all the instruments (see Fig. 7). The threshold of the improvement of spectral smoothness is then defined as $\Delta mbw = (mbw_o - mbw_s)/mbw_o$. Accordingly, an added HRF0 source is considered valid if $\Delta\text{MBW} > \Delta mbw$.

When an F0 hypothesis meets the requirements for a valid estimate, it is removed from the hypothesis list $\mathcal{H}$ and added into the set of the F0 estimates $\mathcal{F}$. During the progressive increase of the polyphony hypothesis m, the algorithm searches for the matched combinations in $\{\mathcal{C}_m\}_{m=1}^{N_{\max}}$. When a matched combination is no longer found, the consolidation process stops. The polyphony is thus inferred along with the estimated F0s. An algorithm flow can be found in [40].

VII. Evaluation

A. Private Evaluation

A systematic method has been proposed to create a polyphonic music database to evaluate the proposed system [42]. In total, 26 pieces have been prepared for the evaluation. The evaluation metrics take into account the estimation of the number of sources [43] and the *overall accuracy* rate is used as the main criterion:

$$\text{Acc} = \frac{N_{\text{corr}}}{N_{\text{corr}} + N_{\text{miss}} + N_{\text{subs}} + N_{\text{inst}}} \tag{13}$$

where N_{corr} denotes the number of correctly estimated notes, N_{miss} denotes the number of missing notes, N_{subs} denotes the number of substitution notes, and N_{inst} denotes the number of insertion notes. Concurrent sources with their F0s related to the same note are regarded as one single source. The system is evaluated on a frame-by-frame basis, and a correct estimate should not deviate from the ground truth by more than 3%.

Tested on the synthesized polyphonic database, the proposed system, which has been submitted for MIREX 2008 [44], is compared to the version submitted for MIREX 2007 [45] (see Fig. 8). The MIREX'07 version is an earlier implementation of the proposed system. It has a slightly different polyphony inference algorithm and it appears to bias low polyphony [46]. That is, it tends to use fewer sources to explain the observed signal and the accuracy in the estimation of high polyphony is not satisfactory. The proposed system uses the presented polyphony inference algorithm, which improves significantly the accuracy for the polyphony higher than 3. The average accuracy rates of the MIREX'07 version and the presented system are 56.56%, and 64.75%, respectively. The presented system has achieved an improvement of 8% in accuracy. However, the estimation for the polyphony higher than five is still to be improved (see Fig. 9).

B. Public Evaluation

The public evaluation results of MIREX 2007 and 2008 for the subtask "frame-by-frame evaluation" are listed in Tables II and III, respectively. The participants are denoted by the team IDs with numeric labels denoting different versions of the submitted systems. The $*$ mark indicates that no temporal continuity is used in the process. The evaluation database for MIREX 2007 is composed of 20 pieces of real recordings and eight pieces of synthesized music; the database for MIREX 2008 is composed of 28 pieces of real recordings and eight pieces of synthesized music. In order to compare the results, it is suggested to use RK as the baseline method because the same version has been submitted for both years [47]. The authors' team ID is CY in 2007 and YRC in 2008. CY and YRC1 refer to the frame-based multiple-F0 estimation system presented, whereas YRC2 further includes a tracking algorithm [21]. In both evaluations, the proposed system CY/YRC and the other two, RK and PI [17], are ranked at top positions, with a significant accuracy

[5]A smoothed out partial is replaced by the amplitude interpolation of its adjacent partials.

TABLE II
RESULTS OF MIREX 2007: MULTIPLE FUNDAMENTAL FREQUENCY ESTIMATION AND TRACKING (FRAME-BY-FRAME EVALUATION)

ID	RK	**CY∗**	ZR	PI1	EV2	CC1∗	SR	EV1	PE1∗	PL∗	CC2∗	KE2	KE1	AC2∗	AC1∗	VE∗
Acc	0.605	**0.589**	0.582	0.580	0.543	0.510	0.484	0.466	0.444	0.394	0.359	0.336	0.327	0.311	0.277	0.145

TABLE III
RESULTS OF MIREX 2008: MULTIPLE FUNDAMENTAL FREQUENCY ESTIMATION AND TRACKING (FRAME-BY-FRAME EVALUATION)

ID	**YRC2**	**YRC1∗**	PI2	RK	PI1	VBB	DRD	CL2∗	EOS	EBD2	EBD1	MG	CL1∗	RFF1	RFF2
Acc	**0.665**	**0.619**	0.618	0.613	0.596	0.540	0.495	0.487	0.467	0.452	0.447	0.427	0.358	0.211	0.183

TABLE IV
RESULTS OF MIREX 2008: MULTIPLE FUNDAMENTAL FREQUENCY ESTIMATION AND TRACKING (NOTE TRANSCRIPTION)

ID	**YRC2**	RK	PI2	PI1	VBB	ZR1	ZR2	ZR3	EOS	EBD2	EBD1	RFF1	RFF2
F-measure (Onset-Offset)	**0.355**	0.337	0.192	0.247	0.197	0.261	0.263	0.278	0.236	0.158	0.176	0.028	0.032
F-measure (Onset Only)	**0.552**	0.614	0.396	0.470	0.521	0.518	0.520	0.530	0.503	0.384	0.417	0.14	0.132

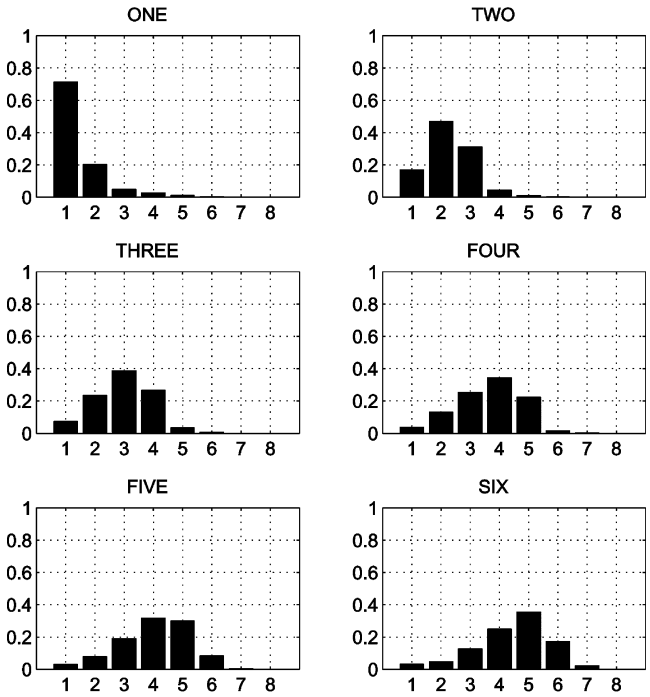


Fig. 9. Distribution of the estimated polyphony for the polyphony from 1 to 6. The title of each subfigure indicates the correct polyphony; the x-axis represents the estimated polyphony; the y-axis represents the percentage of the estimated polyphony among all instances. The peaking at the correct polyphony is observed for the polyphony below five.

gap (5% and above) compared with the rest of the systems. Notice that PI follows the similar scheme as our joint estimation approach, whereas RK is based on an iterative estimation approach [19]. The gain in accuracy appears more significant in the result of the second subtask: note transcription (see Table IV). In this evaluation, a system shall report the onset time, offset time and the average F0 of each note. A total of 30 files were used in this task: 22 real recordings (including six piano solos) and eight pieces of synthesized music. The evaluation criterion is the F-measure [20]. A note is correctly estimated if its F0 does not deviate from the ground truth by more than 3% and the onset/offset time is within $\pm$50 ms range of the ground truth. It is found that the proposed system has a rather precise estimation of the offsets. This result strongly demonstrates the advantage of the adaptive noise estimation and the coherent thresholds derived from musical instrument sounds that allow to detect harmonic sources of relatively weak energy. However, our tracking algorithm does not yet include a probabilistic description of the onsets and is expected to be improved.

The MIREX results contain measured runtime for all algorithms. When comparing runtime of the algorithms, however, one has to take into account that the software means that are used to implement the different algorithms have a significant impact on the runtime. Therefore, the MIREX runtime results have to be treated very cautiously. It is clear, however, that the algorithm presented here will always be relatively costly because for a target polyphony M and N F0 candidates it has to evaluate in the order of $\binom{N}{M}$ possibilities, while the iterative algorithms like RK will test only in the order of NM possibilities. For the detailed list of the participants and the description of their methods, readers are invited to consult the MIREX webpage and the related articles (AC in [10], EV and VBB in [12], KE and EOS in [14], PE in [43], PL in [48], SR in [11], VE and EBB [20], ZR in [15], etc.).

VIII. CONCLUSION AND PERSPECTIVES

We have presented a frame-based multiple-F0 estimation system which analyzes polyphonic music sound signals. The development of the algorithms follows three guiding principles related to the physical properties of harmonic instrument sounds: harmonicity, spectral smoothness, and synchronicity. Several key problems have been treated: noise estimation, harmonic matching adaptive to inharmonic partials, overlapping partial treatment, prevention of subharmonic/super-harmonic errors, the estimation of the number of sources, etc. We have also suggested the derivation of thresholds from musical instrument sound samples, which is coherent for the analysis of polyphonic music signals. The evaluation results demonstrate the competitive performance among the state-of-the-art methods. The proposed F0 estimation system can be improved in two aspects. The joint evaluation part and the polyphony inference part could be combined in an *iterative combination/consolidation* manner. Given a list of F0 candidates, one may iteratively evaluate the validity of an added F0 hypothesis in a hypothetical combination. To develop an efficient and robust algorithm, a strategy to, for each iteration, replace less

likely F0 hypotheses with more probable ones is necessary. The other possibility is to enhance the tracking mechanism.

Most of the research on multiple-F0 estimation aims at using it as the core component within an automatic music transcription system which integrates low-level analyses into a high-level representation as a musical score. The development of an automatic music transcription system requires the integration of the existing music information retrieval (MIR) algorithms such as key estimation, tempo/meter estimation, instrument recognition, etc. In fact, these algorithms may profit from each other to optimize the estimates. For example, an initial guess of the musical instruments can help to extract the underlying F0s. The extracted F0s and the related spectral envelopes can then be used to refine the initial guess of the instruments, which will again help to refine the F0s. Multiple-F0 estimation is especially associated with the following MIR tasks: instrument recognition, melody extraction, key estimation and chord estimation. Other potential applications include the separation/transformation of individual sound sources in a recording, the automatic alignment of a polyphonic recording with a given musical score, etc.

APPENDIX A
HARMONIC MATCHING FOR PARTIAL SELECTION

The source model related to an F0 hypothesis is a set of harmonic grids without specific amplitudes. Given a set of F0 hypotheses, the frequencies and the amplitudes of their partials are to be estimated. The degree of harmonic matching in frequency is evaluated between the model harmonics and the observed peaks. A tolerance interval [49] is designated in the neighborhood of each model harmonic, which allows the inharmonic partials. The spectral peaks situated in the tolerance interval are considered the *matched* peaks, otherwise the *unmatched* ones. For a hypothetical source indexed by m, the *degree of deviation* of the ith observed peak from the hth harmonic is expressed as

$$d_m(i) = \begin{cases} \frac{|f_i - f_{m,h}|}{\alpha_h f_{m,h}}, & \text{if } |f_i - f_{m,h}| < \alpha_h\, f_{m,h} \\ 1, & \text{otherwise.} \end{cases} \tag{14}$$

where f_i is the frequency of the ith observed peak and $f_{m,h}$ is the frequency of the hth harmonic of the model, and α_h determines the tolerance interval $2\alpha_h f_{m,h}$. When an observed peak situates outside the related tolerance interval, it is regarded as unmatched and $d_m(i)$ is set to 1. Therefore, $0 \le d_m(i) \le 1$. Since the partials may deviate from the ideal model harmonics (multiples of F0) due to inharmonicity or frequency modulation, it is necessary to adapt the model harmonics. If the hth harmonic matches the ith peak, the $(h+1)$th harmonic frequency is updated by $f_{m,h+1} = f_i + f_m$, where f_m denotes the F0 value. If the hth harmonic does not match any observed peaks, the $(h+1)$th harmonic frequency is updated by $f_{m,h+1} = f_{m,h} + f_m$. Moreover, since the partials of different sources may fall into one tolerance interval, it is necessary to select the best one for a given F0 hypothesis. The proposed partial selection technique begins with assigning the first partial to the nearest peak. For the consecutive partials, two *peak candidates* are considered: the nearest one, and the one of which the mainlobe covers the related model harmonic. By means of comparing the average amplitude of the previously selected three partials with the amplitudes of the two peak candidates, the peak candidate of a closer amplitude value is selected because it is considered to form a smoother envelope. When no matched peaks are found for a partial, its amplitude is estimated by the interpolation of the neighboring frequency bins around the harmonic frequency.

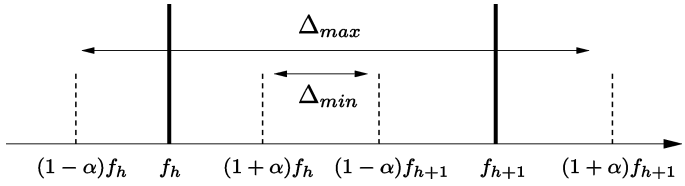


Fig. 10. Allowed frequency differences of two adjacent partials that match to the model harmonics (the two thick vertical lines). The allowed maximum is $\Delta_{\max}$, whereas the allowed minimum is $\Delta_{\min}$. The tolerance interval is defined between the dash lines around a model harmonic.

In the case of monophonic signals, α_h can be set in a way that the tolerance interval equals the F0 to allow inharmonic partials, while prohibiting the overlaps of the tolerance intervals of adjacent harmonics. In the case of polyphonic signals, however, α_h shall be determined in a more precise way to prevent excessive partials of concurrent sources to fall into the same tolerance interval. For a convenient expression, the source index m for the related harmonics is ignored here. Assuming that the values α_h are similar for adjacent partials, i.e., $\alpha_h \approx \alpha$, it is proposed to pose a constraint on the frequency difference of two adjacent partials (see Fig. 10). With the allowed tolerance intervals, the maximum and the minimum of the frequency difference between two adjacent partials are

$$\Delta_{\max} = (1+\alpha)f_{h+1} - (1-\alpha)f_h \approx f_m + (2h+1)\alpha f_m$$
$$\Delta_{\min} = (1-\alpha)f_{h+1} - (1+\alpha)f_h \approx f_m - (2h+1)\alpha f_m$$

in which the approximations $f_{h+1} - f_h \approx f_m$ and $f_{h+1} + f_h \approx (2h+1)f_m$ are used. The allowed frequency difference (half of the tolerance interval) for a peak to match a harmonic is thus $(2h+1)\alpha_h f_m$. Then, α_h can be selected according to

$$\alpha_h(2h+1) \le \beta \tag{15}$$

where β is set 0.3. For the tolerance intervals of lower partials, a minimum constraint is further set according to a quarter-tone frequency resolution. That is, $\alpha_h = 2^{1/24} - 1 = 0.029$ for the first four partials.

APPENDIX B
HYPOTHETICAL OVERLAP TREATMENT

Partial selection estimates the frequencies and amplitudes of the HPS based on harmonic matching. However, the amplitudes of the overlapping partials are ambiguous. Given a combination of hypothetical sources, the positions of the overlapping partials can be easily inferred and the related amplitudes can thus be corrected. The idea is to make the best of the unambiguous information about the non-overlapping partials to remove the ambiguity in the overlapping partials. It is assumed that an overlapping partial still carries important information about at least the HPS that locally has the strongest energy [50]. Therefore, the overlapping partial treatment aims at allocating the overlapping partial amplitude for this HPS. Based on the spectral smoothness

principle, the method to estimate the amplitudes in the overlap positions is described as follows.

- Partials having potential collisions are determined by the peaks that match to more than one hypothetical source. The overlap treatment is carried out in order of the partial frequency.
- In each overlap position, the *local energy* of each HPS is estimated in terms of the interpolation of the amplitudes of the neighboring partials that do not overlap [51]. The amplitude of the overlapping partial is exclusively assigned to the HPS with the largest local energy. The overlapping partial of that HPS is labeled as *credible* and is used like a non-overlapping partial for the consecutive interpolation. This is meant to use as many credible partials as possible for the consecutive overlap treatment. For the rest of the colliding sources, their amplitudes in the overlap position are estimated by the interpolated amplitudes, respectively. This is meant to maintain the local smoothness of the envelope for the partial amplitudes that can not be easily inferred.
- When one of the neighboring partials is overlapped, the amplitude of the non-overlapping one determines the local energy. If both of the neighboring partials are overlapped, the partial for the related source is considered *not credible*. In this case, its amplitude is estimated by the interpolated amplitude if the overlapping partial amplitude is larger than the interpolation.
- When the amplitude of the overlapping partial is smaller than all the interpolated amplitudes of the colliding sources, it is difficult to infer which hypothetical source contributes the most. In this case, the colliding sources share the overlapping partial. The overlapping partial in all HPS is labeled as credible for the consecutive interpolation.

REFERENCES

[1] N. F. Fletcher and T. D. Rossing, *The Physics of Musical Instruments*, 2nd ed. New York: Springer-Verlag, 1998.

[2] J. Harold and A. Conklin, "Generation of partials due to nonlinear mixing in a stringed instrument," *J. Acoust. Soc. Amer.*, vol. 105, pp. 536–545, Jan. 1999.

[3] A. Baskind and A. de Cheveigné, "Pitch-tracking of reverberant sounds, application to spatial description of sound scenes," in *Proc. AES 24th Int. Conf. Multichannel Audio*, 2003, pp. 34–40.

[4] C. Yeh, A. Roebel, and X. Rodet, "Multiple F0 tracking in solo recordings of monodic instruments," in *Proc. AES*, Paris, France, 2006.

[5] A. P. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 804–816, Nov. 2003.

[6] C. Yeh, A. Roebel, and X. Rodet, "Multiple fundamental frequency estimation of polyphonic music signals," in *Proc. IEEE, Int. Conf. Acoust., Speech Signal Process. (ICASSP'05)*, Philadelphia, 2005, vol. 3, pp. III-225–III-228.

[7] A. S. Bregman, *Auditory Scene Analysis*. Cambridge, MA: MIT Press, 1990.

[8] P. J. Walmsley, S. J. Godsill, and P. J. W. Rayner, "Polyphonic pitch tracking using joint Bayesian estimation of multiple frame parameters," in *Proc. IEEE Workshop Applicat. Signal Process.Audio Acoust. (WASPAA'99)*, Mohonk, NY, Oct. 1999, pp. 119–122.

[9] M. Davy, S. Godsill, and J. Idier, "Bayesian analysis of polyphonic western tonal music," *J. Acoust. Soc. Amer.*, vol. 119, no. 4, pp. 2498–2517, Apr. 2006.

[10] A. Cont, "Realtime multiple pitch observation using sparse non-negative constraints," in *Proc. 7th Int. Symp. Music Inf. Retrieval (ISMIR'06)*, Oct. 2006, pp. 206–211.

[11] S. A. Raczynski, N. Ono, and S. Sagayama, "Multipitch analysis with harmonic nonnegative matrix approximation," in *Proc. 8th Int. Symp. Music Inf. Retrieval*, Oct. 2007, pp. 381–386.

[12] E. Vincent, N. Bertin, and R. Badeau, "Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'08)*, Las Vegas, NV, 2008, pp. 109–112.

[13] S. Saito, H. Kameoka, T. Nishimoto, and S. Sagayama, "Specmurt analysis of polyphonic music signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 3, pp. 639–650, Mar. 2008.

[14] H. Kameoka, T. Nishimoto, and S. Sagayama, "A multipitch analyzer based on harmonic temporal structured clustering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 982–994, Mar. 2007.

[15] R. Zhou, "Feature extraction of musical content for automatic music transcription," Ph.D. dissertation, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, 2006.

[16] A. Klapuri, "Multipitch analysis of polyphonic music and speech signals using an auditory model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 2, pp. 255–266, Feb. 2008.

[17] A. Pertusa and J. Iñesta, "Multiple fundamental frequency estimation using gaussian smoothness," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'08)*, Las Vegas, NV, 2008, pp. 105–108.

[18] M. Wu, D. Wang, and G. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 3, pp. 229–241, 2003.

[19] M. Ryynänen and A. Klapuri, "Polyphonic music transcription using note event modeling," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA'05)*, Mohonk, NY, 2005, pp. 319–322.

[20] V. Emiya, R. Badeau, and B. David, "Automatic transcription of piano music based on hmm tracking of jointly-estimated pitches," in *Proc. Eur. Conf. Signal Process. (EUSIPCO'08)*, Lausanne, Switzerland, 2008.

[21] W.-C. Chang, A. W. Su, C. Yeh, A. Roebel, and X. Rodet, "Multiple-F0 tracking based on a high-order HMM model," in *Proc. 11th Int. Conf. Digital Audio Effects (DAFx-08)*, Espoo, Finland, 2008, pp. 379–386.

[22] C. Yeh and A. Roebel, "The expected amplitude of overlapping partials of harmonic sounds," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'09)*, Taipei, Taiwan, 2009, pp. 3169–3172.

[23] R. J. McAulay and T. F. Quatieri, "Speech analysis-synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, no. 4, pp. 744–754, Aug. 1986.

[24] X. J. Serra and J. O. Smith, "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Comput. Music J.*, vol. 14, no. 4, pp. 12–24, 1990.

[25] C. Hory, N. Martin, and A. Chehikian, "Spectrogram segmentation by means of statistical features for non-stationary signal interpretation," *IEEE Trans. Signal Process.*, vol. 50, no. 12, pp. 2915–2925, Dec. 2002.

[26] C. Yeh and A. Roebel, "Adaptive noise level estimation," in *Proc. 9th Int. Conf. Digital Audio Effects (DAFx-06)*, Montreal, Canada, Sep. 18–20, 2006, pp. 145–148.

[27] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions*, 2nd ed. New York: Wiley, 1994.

[28] A. Roebel and M. Zivanovic, "Signal decomposition by means of classification of spectral peaks," in *Proc. Int. Comput. Music Conf. (ICMC'04)*, Miami, FL, 2004, pp. 446–449.

[29] M. Zivanovic, A. Roebel, and X. Rodet, "Adaptive threshold determination for spectral peak classification," *Comput. Music J.*, vol. 32, no. 2, pp. 57–67, 2008.

[30] B. Rivet, L. Girin, and C. Jutten, "Log-Rayleigh distribution: A simple and efficient statistical representation of log-spectral coefficients," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 796–802, Mar. 2007.

[31] A. V. Oppenheim, "Speech analysis-synthesis system based on homomorphic filtering," *J. Acoust. Soc. Amer.*, vol. 45, no. 2, pp. 458–465, Feb. 1969.

[32] Y. Qi and R. E. Hillman, "Temporal and spectral estimations of harmonics-to-noise ratio in human voice signals," *J. Acoust. Soc. Amer.*, vol. 102, no. 1, pp. 537–543, Jul. 1997.

[33] A. Stuart and J. K. Ord, *Kendall's Advanced Theory of Statistics*, 6th ed. New York: Oxford Univ. Press, 1998, vol. 1: Distribution Theory.

[34] W. Hess, *Pitch Determination of Speech Signals*. Berlin, Heidelberg, Germany: Springer-Verlag, 1983.

[35] G. Peeters, "a large set of audio features for sound description (similarity and classification) in the CUIDADO Project," 2003, Tech. Rep.

[36] L. Cohen, *Time-Frequency Analysis*. Englewood Cliffs, NJ: Prentice-Hall, 1995.

[37] A. Roebel, "A new approach to transient processing in the phase vocoder," in *Proc. 6th Int. Conf. Digital Audio Effects (DAFx-03)*, London, U.K., 2003, pp. 344–349.

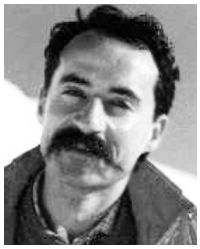
[38] H.-P. Schwefel, *Evolution and Optimum Seeking*. New York: Wiley, 1995.
[39] A. Klapuri, "Multiple fundamental frequency estimation by summing harmonic amplitudes," in *Proc. 7th Int. Conf. Music Inf. Retrieval (ISMIR'06)*, Vienna, Austria, 2006, pp. 216–221.
[40] C. Yeh, "Multiple fundamental frequency estimation of polyphonic recordings," Ph.D. dissertation, Université Paris VI, Paris, France, 2008.
[41] M. Goto, "RWC music database: Music genre database and musical instrument sound database," in *Proc. 4th Int. Conf. Music Inf. Retrieval (ISMIR 2003)*, Baltimore, MD, Oct. 27–30, 2003, pp. 229–230.
[42] C. Yeh, N. Bogaards, and A. Roebel, "Synthesized polyphonic music database with verifiable ground truth for multiple-F0 estimation," in *Proc. 8th Int. Conf. Music Inf. Retrieval (ISMIR'07)*, Vienna, Austria, Sep. 23–27, 2007, pp. 393–398.
[43] G. E. Poliner and D. P. Ellis, "A discriminative model for polyphonic piano transcription," *EURASIP J. Adv. Signal Process.*, 2006.
[44] "Multiple Fundamental Frequency Estimation & Tracking," MIREX, 2008. [Online]. Available: http://www.music-ir.org/mirex/2008/
[45] "Multiple Fundamental Frequency Estimation & Tracking," MIREX, 2007. [Online]. Available: http://www.music-ir.org/mirex/2007/
[46] C. Yeh, "Multiple F0 Estimation for MIREX 2007" The 3rd Music Information Retrieval Evaluation eXchange (MIREX'07), 2007.
[47] M. Ryynänen and A. Klapuri, "Automatic music transcription using note event modeling for MIREX 2008," The 4th Music Information Retrieval Evaluation eXchange (MIREX'08), 2008.
[48] P. Leveau, E. Vincent, G. Richard, and L. Daudet, "Instrument-specific harmonic atoms for mid-level music representation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 116–128, Jan. 2008.
[49] T. V. Sreenivas and P. V. S. Rao, "Functional demarcation of pitch," *Signal Process.*, vol. 3, pp. 277–284, 1981.
[50] S. T. Roweis, "Factorial models and refiltering for speech separation and denoising," in *Proc. 8th Eur. Conf. Speech Commun. Technol.*, Geneva, Switzerland, 2003, pp. 1006–1012.
[51] R. C. Maher, "Evaluation of a method for separating digitized duet signals," *J. Audio Eng. Soc.*, vol. 38, no. 12, pp. 956–979, Dec. 1990.

Chunghsin Yeh (M'09) received the B.S. degree in mechanical engineering and the M.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan in 1998 and 2000, respectively, and the M.S. ATIAM (acoustics, signal processing, and computer science applied to music) degree and the Ph.D. degree in computer science from University Paris 6, Paris, France, in 2003 and 2008, respectively.

In 2008, he joined the analysis–synthesis team of IRCAM, Paris, for research and development. His research interest focuses on music signal analysis and processing.

Axel Roebel (M'08) received the Diploma in electrical engineering from Hannover University, Hannover, Germany, in 1990 and the Ph.D. degree (*summa cum laude*) in computer science from the Technical University of Berlin, Berlin, Germany, in 1993.

In 1994, he joined the German National Research Center for Information Technology (GMDFirst), Berlin, where he continued his research on adaptive modeling of time series of nonlinear dynamical systems. In 1996, he became an Assistant Professor for digital signal processing in the Communication Science Department, Technical University of Berlin. In 2000, he obtained a research scholarship at CCRMA Standford University, Stanford, CA, during which he worked on adaptive sinusoidal modeling, and in the same year he joined IRCAM, Paris, France, for working on the analysis–synthesis team doing research on frequency-domain signal processing. In summer 2006, he was Edgar–Varse Guest Professor for computer music at the electronic studio of the Technical University of Berlin. Since 2008, he has been Deputy Head of the analysis–synthesis team of IRCAM. His current research interests are related to music and speech signal modeling and transformation.



Xavier Rodet's (M'06) is currently with the Institut de Recherche et Coordination Acoustique/Musique (IRCAM), Paris, France. His research interests are in the areas of signal and pattern analysis, recognition, and synthesis. He has been working particularly on digital signal processing for speech, speech and singing voice synthesis, and automatic speech recognition. Computer music is his other main domain of interest. He has been working on understanding spectrotemporal patterns of musical sounds and on synthesis-by-rules. He has been developing new methods, programs, and patents for musical sound signal analysis, synthesis, and control. He is also working on physical models of musical instruments and nonlinear dynamical systems applied to sound signal synthesis.